Document servers and the electronic distribution of grey literature

van Horck, A.J.M.; Tuck, B.

Publication date:
1994

Document Version
Publisher's PDF, also known as Version of record

[redacted]

ITK RESEARCH REPORT

INSTITUTE FOR LANGUAGE TECHNOLOGY AND ARTIFICIAL INTELLIGENCE INSTITUUT VOOR TAAL- EN KENNISTECHNOLOGIE

ITK Research Report No. 55

Document Servers and the Electronic Distribution of Grey Literature

Arthur van Horck

ITK
Warandelaan 2
P.O. Box 90153
5000 LE TILBURG
Harry.Bunt@kub.nl

Bill Tuck

3 Thornhill Square
London, N1 1BQ
United Kingdom

Bill.Tuck@uk.bl.london

November 1994

ISSN 0924-7807

Document Servers and the Electronic Distribution of Grey Literature: Strategies for Further Development*

Arthur van Horck
ITK, Tilburg University
P.O. Box 90153, 5000 LE Tilburg
The Netherlands
Phone: +31 13 662232
Fax: +31 13 662537
E-mail: Arthur.vanHorck@kub.nl

Bill Tuck
3 Thornhill Square
London, N1 1BQ
United Kingdom
Phone: +44 71 700 4293
Fax: +44 71 700 5877
E-mail: Bill.Tuck@uk.bl.london

May, 1994

"If we do not succeed,
then we face the risk of failure"
Former US Vice President Dan Quayle

*This report was commissioned by Stichting Centrum voor Bibliotheekautomatisering PICA and The British Library

Abstract

This report provides an overview of some of the technical and administrative issues connected with the problem of integrating document servers (containing holdings of articles in electronic form) with standard document delivery systems.

It also describes the prerequisites for a pilot project for the (electronic) dissemination of so-called Grey Literature. The notion of Grey Literature is expanded, and the foundations of means of capture, search and delivery are laid out.

STATUS:
Final Version 1.5
September 1994

Contents

Glossary

API	Applications Programming Interface — general descriptive term
Archie	
Ariel	
ARTTell	Article Request Transmission (by Telex) — a British Library document supply service
ARTX.400	Article Request by X.400 — a British Library development
ATM	Asynchronous Transfer Mode — a high-speed switch technology
Attent	Tilburg University's Grey Literature bibliographical database
CCITT	Consultative Committee on International Telephone and Telegraph — the standards body of the International Telecommunications Union
DDX	Document Delivery by X.400 — A British Library project
BL DSC	
DTD	Document Type Description — a component of SGML
EDIL	Electronic Document Interchange between Libraries — a research and development project under the EC Libraries Programme
Foudre	
ETI	Electronic Text Initiative
FTP	File Transfer Protocol — an Internet standard
FTAM	File Transfer, Access & Management — an ISO standard
GEDI	Group for Electronic Document Interchange
GGC	Gemeenschappelike Geautomatiseerde Catalogus — Online Shared Catalogue
GIF	Graphical interchange format
Gopher	Menu based, strongly hierarchical, Internet information service
HTML	Hypertext Markup Language — a DTD
HTTP	Hypertext Transfer Protocol. A Stateless Search, Retrieve and Manipulation Protocol
ISO	International Organisation for Standardisation
IETF	Internet Engineering Task Force — standards body for the Internet
IAFA	Internet Anonymous FTP Archives — a working group
ILL	Inter-Library Loan Protocol — an ISO standard
MIME	Multimedia Internet Mail Extensions — an Internet standard

MPEG	Moving Pictures Experts Group — an ISO draft standard
OBN	Open Bibliotheek Netwerk, open library network
ODA	Open Document Architecture — an ISO standard
OPC	Online Publieks Catalogus — Online Public access catalogue
OSI	Open Systems Interconnection
PDL	Page Description Language
RAPDOC	Pica network-based document delivery system
RFC	Request for Comments — Internet discussion papers & standards
RTF	Rich Text Format — a Microsoft standard
SGML	Standard Generalised Markup Language — an ISO standard
SMDS	Switched Multimegabit Data Service
SMTP	Simple Mail Transfer Protocol — an Internet standard
SR	Search & Retrieve Protocol — an ISO standard
TCP/IP	Transmission Control Protocol / Internet Protocol — the standard protocol used on the Internet)
TEI	Text Encoding Initiative
TIFF	Tagged Image File Format — an Adobe Corp. standard
URL	Universal Resource Locator — an Internet standard
URN	Uniform Resource Name — an Internet standard to be
WAIS	Wide Area Information Server
WWW	World Wide Web
Z39.50	the US ANSI version of the SR protocol

Introduction

The following study is intended as the first phase of a project to develop a number of demonstrator systems of *document servers*. As such it is primarily concerned to give a general overview of the field and to ask the right questions. A second phase will look into the question of detailed mapping of the model that is expected to be derived from the first phase to specific hardware and software and development of prototype document servers to beimplemented for user trials and evaluation.

The growth of network databanks, document archives, and access tools (WAIS, Gopher, WorldWideWeb, etc.), on the one hand, and the increasing demand for library-based document request and delivery services (RAPDOC, Ariel, ARTTel), on the other, raises the question of how these two rather different approaches might be reconciled. It would be unfortunate if the two worlds developed along mutually incompatible lines.

The initial phase of the study therefore looks at the question of the relationship between current library-based information services (particularly document delivery) and the growing range of network-based information services. Can the general *document request and delivery* model be extended to include the kind of services currently provided by **Anonymous FTP**? On the other side, could the emerging model of *linked hypertext access to network resources* (typified by WorldWideWeb or WWW) be extended to include document request and delivery services of the more conventional kind?

These two approaches to the problem form complementary perspectives and the general objective of this study is to see whether some reconciliation of these two points of view is possible.

In parallel with this general question, however, a specific concern of the project is with the dissemination of so-called grey literature through electronic means. This addresses the very practical issue of how best to set up a document server for a particular kind of document – the research memoranda generated by an academic organisation.

This gives rise to the appropriately-named *Grey Files project*, which has as its particular objective the setting up of an experimental document server for material generated in this way – to provide an electronic means of informal (or semi-formal) academic publication.

How the Study Was Carried Out

Much of the background information for the report has come from the authors' experience over a number of years in working with electronic document delivery systems and with various forms of network publishing.

In addition, a limited survey was carried out of a number of document server operations available to the authors. These mostly consisted of document archives on the Internet, but covered a range of technologies from WAIS to Gopher, WorldWideWeb and email. Most of these are currently operating on a limited experimental basis, but the intention of many is to move towards fully supported (and even commercial) services in the near future.

Finally, discussions with librarians, publishers, indexing agents and others closely involved in the serials publishing world have served to provide background information and insight into the kinds of response that such developments in electronic publishing are likely to evoke.

Structure of the Report

After a short introductory chapter on Document Origination and Document Types, each of the following chapters of the report first discusses the general question of access to electronic documents and then considers the particular issues as they relate to Grey Files of the kind defined above (ie. academic research memoranda). Chapter 2 defines, in detail, what Grey Files are, from the point of view of this project and contrasts the proposed project with two other Dutch projects with similar aims.

Chapter 3 (Encoding Formats) considers document encoding formats, in general, then looks at what kinds of files are to be expected on the server and the particular requirements introduced by Grey Files.

Chapter 4 (Information Retrieval and Document Request) deals with the problems of finding and requesting electronic information. It also deals with the question of how Grey Files are produced and catalogued. Another section is devoted to a discussion of three network navigation methods: Gopher, WAIS and the World Wide Web. These are again contrasted in Chapter 5.

Next, Chapter 6 (Printing and Display) deals with the issues of transfer and display of Grey Files. As the Grey Files project is primarily concerned with electronic dissemination of grey literature, hardcopy distribution is addressed only as an aside. Chapter 7 provides a general summary of the report. Finally, Chapter 8 provides a number of recommendations for work to be carried out.

1. Document Origination and Document Types

This section is concerned with the question of where the documents come from. It will also consider the kind of document and its general characteristics, though not its specific format. Several examples of contrasting document databases are presented.

For the purposes of this study, specific examples are taken from two broad areas: Research memoranda (including research theses) and standards documents (including CCITT/ISO specifications, Internet RFCs or IETF documents). A third category that could be considered is historical material. In all cases, an important consideration will be whether the material lies outside of copyright restrictions.

Among the examples of research memoranda archives that have been considered in the UK, a recent experiment between University College London (UCL) and Brunel University investigated the use of the X.500 Directory Services Protocol as a means of indexing and accessing a database of machine-readable documents. The documents themselves had been created within the Computer Science Departments of the two universitites, using a wide variety of text-processing systems.

One of the significant problems in creating such an archive is to find a satisfactory way of dealing with a multiplicity of formats. Even within a single university department there is a high probability that many different formats will arise – from the simplest unformatted ASCII text to complex PostScript documents. Some possible techniques for dealing with this problem are discussed in the next section.

Archives of research memoranda are beginning to appear on the Internet. Access is at present rather cumbersome and is generally based on Anonymous FTP or an email `SEND <filename>` request to the appropriate mail-server. Informal, if ad hoc, standards have evolved in both cases which provide a practical (if sometimes cumbersome) document request and delivery service. Techniques by which this might be developed into a more fully-featured service are discussed later in this report.

One example of such an archive is that held at Imperial College London. In addition to software, this holds a considerable mass of documentation material, including a substantial portion of the CCITT standards archive (the so-called "coloured books"). Such material could be useful as a test database of documents. The problem is to find a satisfactory way of indexing the material – the conventional technique of modeling the database simply on the Unix file structure is extremely unwieldy and totally opaque to the uninitiated or non-Unix user.

Closer to the *research memoranda* idea is the document database operated at the University of Newcastle in the UK. This is called *Mailbase* and is accessed primarily by the method of email request outlined above. The format for a request is a message containing just the text: **`SEND <list-name> <filename>`** where `<list-name>` is the name of the *field of interest* (in fact a mailing list for a special-interest group) and `<filename>` is the (unix-style) name of the document itself. A list of available documents can be retrieved as an index document – no online

index searching is possible. This email-based document server is typical of many operating on the Internet.

One alternative approach is to treat the document database as a homogenous collection of similarly encoded items. An example of this is the CORE database of ACS journals. CORE (Chemistry Online Research Experiment) took as its source material the 24 journals of the American Chemical Society and encoded them digitally, by matching up the original typesetting tapes with the scanned images of the printed pages (in order to include the pictorial information that did not appear in the typsetting tapes). The entire collection of several thousand articles has been retrospectively encoded in SGML and now forms a single monolithic database, accessible through *hypertext* links between articles as well as by more conventional key-word searching techniques. In this sense, it is not a *document delivery* system so much as a *page delivery* system. One consequence is that the bandwidth required on the communications channel between server and user's screen is very high – at least up to ethernet LAN speeds, if not higher. This makes it unsuitable for running on a wide area network (at least with the speeds generally available at present).

What the CORE project serves to illustrate is the difference between a *document server* and a *page server*. With the growth of network access to electronic document archives offering instantaneous delivery on-demand (with a response time limited to a large extent only by the speed of the network) it becomes possible to consider a mode of working in which many documents from many different sources might be retrieved at one time for local browsing.

Transferring several hundred documents to a local page server for further processing (rather than to the individual's workstation) could be an efficient way of providing access to a wide range of material. There is every reason to believe that storage and transmission costs will continue to fall, with the effect that it will become increasingly viable to build large databases for this kind of material. Making such resources available over wide area networks, such as Internet, will be a challenge, both technically and administratively.

2. 'Grey Files' Publishing

With the growing adoption of word processing, documents are increasingly likely to have been originated in electronic form – even where they may be printed and published in hard copy. It is tempting, therefore, to envisage the collection of such documents into databases of machine-readable material which may be searched and otherwise processed by users, with online access via computer networks.

Such collections of *grey literature* could constitute a valuable resource of information, particularly if it were possible to organise it in ways that made for easy identification of relevant material and access to original documents. The implementation of such a system is the primary objective of the Grey Files project. The motivation for the project comes, in part, from a desire to see if current methods of handling document search, retrieval and delivery can be extended to accommodate this new domain.

2.1. Definition

The Grey Files project is specifically concerned with a particular subset of the so-called grey literature. As the name implies, we are dealing here with documents – be they textual or graphical – that are available in electronic, machine-readable form. But not all electronic documents qualify as Grey Files. For an electronic document to be considered as a grey file, a number of criteria have to be met:

- The document must be free from (legal) copyright;
- The document must have been produced in an academic setting – normally in a series of similar documents produced and published by some academic institution;
- The document must – usually – be reporting on (details of) work in progress; this means that, strictly speaking, theses are out as far as this project is concerned.
- The document may have been – or may be – submitted to conferences, which means that it may also be available in hardcopy form, as a conventional document (from a librarian's point of view).

Even with such a restricted definition, it would be useful to bear in mind both the potential complexity and the scale of the problem. Nor is it without contradictions.

An organisation (for example, a university) may generate perhaps ten times as many research memoranda as ever appear in print through the standard publishing channels. While such documents may be of value internally to the organisation, it is less clear that this value extends beyond, or even that the organisation would wish them to be openly available. In any case, those that do have worth beyond the institution would normally be published through the conventional channels. Direct electronic publication, in other words, could lead to a bypass of the quality control and disclosure constraints of the traditional publishing system. It is for this reason that the Grey Files project only concerns itself with official, institutionally sanctioned reports and series.

It is questionable, for a number of reasons, whether a national organisation (such as the British Library) should be interested in maintaining sources of electronic documents of this kind (ie. research memoranda or grey literature – as opposed to electronic versions of conventionally published texts or historic manuscripts). One model suggested for Grey Files publishing views it as a set of document servers, one at each participating institution.

In this sense it represents, in principle, an inherently distributed document database rather than a centralised one (in contrast, for example, to the British Library Document Supply Centre in the UK). Nevertheless, in the Netherlands, the Royal Library is formulating a strategy towards its functioning as a legal repository for electronic documents. It is not yet clear how such a policy might affect Grey Files.

The arguments above also suggest that a certain amount of discretion might be appropriate in selecting suitable material for *publishing* in this form. It should, for example, fulfill certain conditions of quality and bibliographic control. Examples might have included the accepted theses of the institution (though, at least in the case of the UK, it is not clear that copyright restrictions would actually permit this).

In addition, if electronic publication is to be organisation-based rather than publisher-based, then it is likely that the natural focus for such activity should be the library, at least in the case of universities. This is a new role which libraries themselves have begun to explore – essentially that of helping academics towards electronic publication. Libraries are beginning to offer their services to the institution as a means for creating online archives of documents [Tuck, (1992)]. At the same time, cooperation between libraries should ensure that similarly published material could be made available between organisations.

Much of the underlying motivation for the present study lies in the belief that this model offers an opportunity for the library to regain the initiative, as primary information source, from both the commercial publishing world and the burgeoning world of informal network publishing. Realistically, some form of partnership between all three domains would be the most desirable arrangement.

2.2. Relationship to Other Projects in The Netherlands

There are currently three other official projects in this area in The Netherlands:

SURFdoc – in Groningen and Leiden is the SURF pendant of Grey Files. Two separate projects have been defined, and are currently under way:

- Eldoradoc – Rijks Universiteit Groningen. As with Grey Files, this project also aims at integration with the Pica infrastructure, but has as its main object doctoral theses. The delivery vehicle of choice for disclosure is a Gopher/WAIS combination. A later version is expecting to use MIME email for document delivery.
- RUL Research reader – Rijks Universiteit Leiden is aimed at disclosure through Wais and Gopher. Several series of in-house publications will be made available, either in full text or in the form of scanned images.

KUN – Nijmegen University, is concerned with doctoral theses only

Images Prentenkabinet – at Leiden University (No further information is at present to hand on these last two projects)

2.3. Relationship to Document Delivery

The basic aims of the projects listed above are the same: to make available, electronically, materials (such as doctoral theses) that are already available in another form through traditional channels. They differ from the Pica RAPDOC project, in that their primary aim is to make the documents accessable electronically (ie.

for on-screen display), whereas in the RAPDOC project final delivery is on paper and only the transmission is electronic.

In RAPDOC (as in EDIL, ARTTel, FOUDRE and many similar document delivery systems) the primary information consists of published articles held in the serials collections of the participating libraries. Encoded as scanned images (usually in TIFF format), these can be transmitted on request across the network. On receipt they will generally be printed out.

The Grey Files proposal, on the other hand, is concerned with material that is already in machine-readable electronic form. The basic question that the present report is trying to address is whether access to this material can be provided by the same mechanisms used in RAPDOC. Comparable (and compatible) standards of cataloguing and bibliographic control will clearly be important.

3. Document Encoding Formats

What formats will documents be in? Research memoranda generated within the usual anarchic domain of a university research department will inevitably use a wide variety of incompatible document encoding and formatting schemes. Although there is no shortage of standard encoding formats available, there is a problem in getting authors to adhere to them. The same applies to the bibliographic control and cataloguing elements. In conventional serials publishing these operations are delegated to others, such as publishing and abstracting services. How would it work electronically?

3.1. General Document Encoding Formats

Attempts to enforce standards at the author level, by encouraging the use (by authors) of OSI document standards (SGML or ODA), for example, have been largely ineffectual. Nor is there reason for much optimism that any such standardisation could ever be successfully imposed.

The alternative is to devise a way of indicating the encoding format of the document itself. This is possibly a more promising approach, but as yet remains poorly developed. For example, on the Internet document servers it is customary to indicate the type of file by conventions within the filename – if it terminates with a `.txt` it is an unformatted plain text file, while `.ps` indicates PostScript. Such a method has obvious limitations – particularly when trying to deal with different forms of compression. Furthermore, the problems can only increase as new and more complex forms of data become available through archives, such as bitmaps, colour images, sound, etc.

Clearly, what is needed is a descriptor file that may be appended (or prepended) to the document to indicate the form of encoding, together with any other information that may be relevant, such as a set of bibliographic descriptors.

For conventional documents, the GEDI initiative has gone furthest in this direction by devising a set of tags that enable a document to at least be identified with a high degree of precision. Although it is primarily targetted at the problem of

specifying articles published in the traditional serials literature, encoded as TIFF images, it could be adapted to a much wider range of material [GEDI, (1991)].

In general, the question of document formats is a complex one and is unlikely to meet with a simple solution. In some respects, the document image (rather than machine readable text) as a TIFF file forms the lowest common denominator, as well as the most easily generated format able to include the majority of representational forms – text, graphics and images. Its disadvantage is bulk and the absence of machine searchability. With increasing network bandwidths, storage capacities and processor power, these disadvantages will retreat. For this reason, one possible option (in the context of the present study) might have been to concentrate on the particular problems of providing access to material encoded in this format. In reality, it would seem to make little sense to go to the additional work of rescanning material that is already available in electronic form. For retrospective conversion, however, it may be the only method available.

The British Library Document Supply Centre, for example, is presently engaged in a project to store the contents of 50 journals in scanned-image form. This will provide a test database of a substantial size from which documents may be delivered on request. The experiment should serve to ellucidate some of the more complex questions of document delivery from an online archive, such as the overall cost per request – given that useage rates for such material are relatively low (less than 10% of articles are requested more than once) the costs of input, indexing, storage and delivery can seem very high compared to more traditional methods.

Another reason that could be used to justify concentrating on image-encoded documents is that it fits with the present inter-library-loan services. Serials holdings are almost exclusively in paper form. This is the kind of material that document supply libraries know how to manage. And the only effective way of transmitting paper-to-paper electronically is by encoding it as a fax (or TIFF) image. Even in cases where publishers, with access to machine readable tapes, have sought to produce electronic versions of their titles (such as with the ADONIS project), it has proved easier to rescan than to derive a uniform encoding from the original tapes.

3.2. Document Encoding Formats for Grey Files

Documents that would qualify for inclusion in the Grey Files databank are currently produced using a number of different text processors, with different electronic output formats. One possibility, as suggested above, would be to translate all of these to one single format, in order to guarantee ease of use, both for the end user as for the system maintainer. For the end user, because he/she would be confronted with one single format, and for the system maintainer because of the fact that systems requirements will be easier to foresee.

One problem with this is that conversion of a document into another format may cause problems in the area of integrity: is a document still the same, after such a conversion?

3.3. Document Production and Version Control

If format conversion takes place, which version of the document will count as the original? In conventional publishing it is the printed form of the actual publication that takes precedence, not the author's original manuscript. In the electronic case, it could be the library that plays the role of publisher, responsible for the authenticity and integrity of the document database.

Once entered into the database, the document should not be altered. Nevertheless, it may be reasonable to anticipate that the author would wish for the opportunity to make corrections or amendments at any stage after initial *publication*. While this may be technically feasible, it is likely to prove undesirable. There is a risk that the document might appear in a number of locations, all of which would need to be simultaneously upgraded, and the responsibility for this level of version control could be heavy.

It may be better to think of each document having only one authenticating record – only one database on which it can be assumed to be the *correct* version. This may not be a very satisfactory solution, but it is difficult to envisage a more realistic alternative.

While an alternative model, based on the kind of version control methods used in software development and software publishing, might be possible, it could prove unwieldy. Documents are not really like software, in that there is little justification for continual updates. The cataloguing of electronic archives (documents as well as software) on the Internet is the subject of considerable debate at the present moment. (See also the E-doc project of the Netherlands Royal Library.)

4. Information Retrieval and Document Request

Returning to the basic question: How will material be found and how will requests for that material be handled?

4.1. General Issues

Finding material on the Internet is not always easy. Unlike the formal publication process, where journals adhere to strong traditions of what is considered to be good practise in providing the information necessary for identifying material, the Internet is anarchic. There is little that could reasonably be compared with the abstracting services or the creation and management of online bibliographic databases, with respect to the electronic publications on the Internet. This partly reflects the fact that network publishing is not *commercial* in any easily recognisable sense.

Despite this, over recent years there has been a remarkable growth in the range of tools available for resource location on the network. With increasing sophistication their effectiveness will undoubtedly grow. Nevertheless, it is difficult to imagine that the Internet (including its European offshoots, the academic networks), as a global information resource, will ever become as well organised as,

say, our national library systems. This is partly due to the way the Internet itself originated and is managed – it has frequently been described as "the world's largest functioning anarchy". This does not preclude, of course, that within certain sub-domains coherent systems for information access can be organised. In fact, both RAPDOC and ARTTel (particularly in its X.400 extensions) are intended to operate across Internet connections.

Part of the difficulty with present access tools may be that they try to be all-inclusive, with little (if any) differentiation between software source files and document files, for example. Another factor is that the source information is poorly indexed. Archie servers, for example, may collect filename information from several thousand databases and will generally hold several million records, but the usefulness of this information is severely limited by the fact that it consists of just filenames, rather than anything that might correspond to a bibliographic record. In principle, it need not be like this. Archive managers are encouraged to create index files of their holdings with keyword references. Archie then assembles these into a grand index accessible via the `whatis` command. The response to a query of the form `whatis <word>` is a list of filenames recorded as having that keyword reference. The user then has to locate the whereabouts of the file by doing a standard `prog` Archie request on the filename. It works, but the results are far from impressive.

The main problem is poor indexing. There is little real incentive for an archive manager to provide detailed indexes to the files within his database. The second problem (at least from personal experience) is that system response time can be very bad. This probably depends on the loading level of the Archie server itself, or the network bandwidth. Either way, there may be little real incentive to improve this since the service is free of charge to the user. A more important criticism, however, is that it does not seem to be the most rational way to build an index and access file – a single search should produce both the name of the file containing the keywords and its location (the analogy with *Inside Information* is that the shelf location mark at BL DSC is stored along with the bibliographic record).

WAIS takes a different approach by allowing full-text searching on file contents, but it is limited by the need for the user to know in which dataset to direct the search. Given that the number of datasets available from a given WAIS server may number several hundred this can itself create a significant problem.

4.2. Cataloguing in The Netherlands' Grey Files Projects

There is, as yet, not very much coordination between the different projects listed in the previous chapter, although the Royal Library and Pica are currently trying to get the projects on a single track by

- coordinating format definitions for Pica's Online Shared Catalogue (GGC)
- giving guidelines for cataloguing the materials, and
- coordinating the interaction with the GGC

The debate on cataloguing electronic archives has thus been settled in the Netherlands by a general consesus to adopt the proposed Royal Library description format for the central Pica database.

It is felt that cataloguing should be done in the Intelligent Bibliographic Workstation, along the lines that are currently being followed for traditional publications: the library concurrently updates the On-Line Shared Cataloguing System (GGC); the On-Line Public Access Catalogue (OPC) is subsequently updated from the GGC. It may also be possible to extend or adapt the current special interest databases (ATTENT, for instance) for use over the network, and in accordance with what is laid out in the E-doc subproject on data structures

Eventually, Grey Files will thus integrate seamlessly in the Open Library Network (OBN), by its being catalogued and indexed in the common Pica databases, by its being accessible through the same technical infrastructure, as well as being accessible from the "outside".

4.3. Gopher, WAIS and WorldWideWeb

The main information retrieval systems in use on the Internet are Gopher, WAIS and the World Wide Web (WWW). All work on the client-server paradigm, and all provide some degree of support for multimedia data. Considering that the Internet is a conglomerate of mainly academic institutions, where users of different levels of sophistication, with very different needs with regard to hardware and software will mix and communicate, obvious requirements for navigation tools (in no particular order) might be:

- Support for multiple platforms and for a wide range of media types through external viewers (this is cared for in all three systems)
- Support for sophisticated presentation and synchronisation requirements (unfortunately, this is not taken care of in these systems)
- Support for database querying. (This tends to be limited to "keyword" searches, but current developments (such as adoption of Z39.50 in Gopher and the WWW) will make more sophisticated queries possible in the near future.)
- Adequate response times. (These vary considerably for all three systems, depending on the network distance between client and server, as well as the time of day and the available network bandwidth.)
- Support for user authentication, monitoring and billing. (There is currently little in this respect available, although these are planned for the WWW.)
- Support for hyperlinks (only in WWW).
- Support for user annotation of documents. (Some, but not all, clients support this.)
- Support for a mail responder (only in WWW).

- The only system sufficiently complex to warrant an authoring tool is the WWW, which has editors to support its hypertext markup language.

As for all these on-line provision services, maintaining quality of *service* is the responsibility of the body that provides the service. There is no reason to suspect that either one of WWW or Gopher should be more easily maintained than the other, as long as care is taken to surround the service with an adequate organisation.

4.4. Gopher

Gopher presents the user with a hierarchical arrangement of nodes which are either directories (menus), leaf nodes (documents containing either text or other media types), or search nodes (allowing some set of documents to be searched using keywords, possibly using WAIS). A range of media types is supported. Extensions currently being developed for Gopher (Gopher+) provide better support for multimedia data. Gopher has a very high penetration (there are over 1000 Gopher servers on the Internet), but it does not provide hyperlinks and is extremely hierarchical.

4.5. WAIS

Wide-Area Information Servers (WAIS) allow users to search for documents in remote databases. Full-text indexing of the databases allows all documents containing particular (combinations of) words to be identified and retrieved. The client talks to the server using an extended version of the Z39.50 protocol. This is now aligned with the ISO Search and Retrieval protocol for library applications. Note, however, that the WAIS Z39.50 implementation does not support full boolean operations, and neither supports the facilities for access control and charging. This may change in the near future, when WAIS adopts the now current Z39.50 version 2.[1] Non-text data (principally image data) can be handled, but indexing such documents is only performed on the document file name, severely limiting its usefulness. However, WAIS is ideally suited to text search applications.

4.6. The World Wide Web

The World Wide Web is a large scale distributed hypermedia system. The Web consists of nodes (also called documents) and links. Links are connections between documents. To follow a link, a user clicks on a highlighted word in the document (or types a number in a menu) which causes the linked-to document to be retrieved and displayed. A document can be one of many media types, or it can be a search node in a similar sense to Gopher. Any file available for anonymous FTP can be immediately linked into the Web by making a single link to a document already in the Web, and running a local FTP or HTTP daemon. The WWW addressing

[1] As of February 16, 1994, freeWAIS-version 1.0, which has been tested under SunOS, Ultrix, and Microsoft Windows, now provides Z39.50 Version 2 support. (http://cnidr.org/cnidr_projects/freewais1.0.tech-spec.html)

method means that WAIS and Gopher servers may also be accessed from (indeed, form part of) the Web. The World Wide Web has a smaller penetration than Gopher, but is growing faster. The Web technology is currently being revised to take better account of multimedia information.

Browser software is available for a large number of systems, including line-mode dumb terminals, `curses` library support, Macintosh, X/Motif, X11, PC/MS Windows, Next. Server software is available for Unix, Macintosh, as well as VM mainframes.

4.7. URLs

All nodes on the Web are addressed using the Universal Resource Locator (URL). A URL is a name for an object, which may be a document or an index, on the Internet. It has the general form

```
<scheme> : <path> [ # <anchorid> ]
```

The `<scheme>` identifies an access protocol or method for the object. The `<path>` component locates the document in a way that makes sense for the access method. The optional `<anchorid>` is used for addressing within an object. Its interpretation is not defined in the URL specification.

4.8. Mosaic

The Mosaic project, located at the US National Centre for Supercomputing Applications (NCSA) is developing a networked information system intended for wide area distributed asynchronous collaboration and hypermedia-based information discovery and retrieval. Mosaic is specifically aimed at scientific research workers, and has adopted the World Wide Web as the core of the system.

4.9. Extensions of the Archie Model

Another approach to imposing a coherent structure upon a distributed model might be to adopt some features of Archie, WAIS and the rest, and to combine these with the more traditional approach of library-based search tools such as CARL's *Uncover*, BL's *Inside Information* or Pica's RAPDOC.

Inside Information, for example, holds records of the contents of some 10,000 serials titles. These can be searched (either from CD-ROM or online) under author, title, and keyword, in the usual way. Once located, an item can be requested for delivery from the supply library – in this case the BL DSC. Issuing a request simply entails marking the item – the creation and transmission of an ARTTel message is then automatic.

If electronic archives of documents were indexed in a similar way (author, title, keywords and abstract) and if a standard for index formats could be agreed by participating organisations, then a facility like Archie could assemble a corresponding *union catalogue*. This cumulative index would be updated automatically at regular intervals (as Archie does). In this way a substantial archive of online

documents could be assembled from a potentially very large number of individual servers – one for each university or participating organisation, for example. Document ordering would follow a similar process to that of *Inside Information* or RAPDOC, in which flagging an item in the index database will generate an automatic email request for the corresponding source document to be transmitted (again by email, or file transfer).

Clearly, a system of this kind will only work if standard formats for both index record and source document can be adopted. The GEDI consortium has gone some way towards such a specification and may provide a suitable model around which such a system could be developed. GEDI itself is targeted at electronic delivery of printable documents, such as the scanned page images of journal articles. There is no intrinsic reason why it could not be adapted to a wider range of document types, such as PostScript or Acrobat, or even to objects with no print-on-paper equivalent. Neverthless, a strategy of restricting documents to a limited range of format types could be advisable, at least in the initial stages.

The IETF (Internet Engineering Task Force) has itself set up a group to investigate ways of rationalising access to network resources, and network publishing in particular. This goes under the name of the Internet Anonymous FTP Archives (IAFA) working group [Deutsch and Emtage]. The model adopted is very much like the *extended Archie* system described earlier. Each archive must include an index in a standard format of all the material in the archive. These indexes will be polled regularly by the *document servers* (corresponding to the present *Archie* servers) to create a kind of union catalogue. In this way the document server provides access to the full distributed database of documents.

Managing document requests in a system of this kind does not appear to be a particularly difficult problem. In the *extended Archie* model, the central index holds records which indicate both the bibliographic data relating to the document and the addressing data of the source from which it came. This should be sufficient to generate an automatic request message once an item/record has been flagged. The address, for example, might consist of the source file-name and the Internet address of the host – sufficient for an automatic FTP, or email transaction. In fact, this already exists in the form of the URL (Universal Resource Locator) as described earlier.

A more complex issue is the implementation of a charging mechanism for document requests. This would be essential for any full-scale commercially supported service. The most obvious way to organise payment would be through centralised billing, presumably running as an adjunct to the indexing and request service. Reconciliation of payment with the various *suppliers* (ie. document archive managers) would be the responsibility of the central authority.

Finally, what would be entailed in extending the *Inside Information* service to cover electronically stored documents held on remote databases? Entries in the index would need to include source file-names and a location code (ie. the URL), for items held electronically (as well as the standard bibliographic data). Request messages would pass through the normal accounting procedures (such as ARP, in the case of BLDSC) and then be forwarded to the remote site, if necessary (perhaps via the EDIL relay). On return, the document would follow the same

path, generating a *shipped* message at the EDIL relay.

5. Document Delivery Methods

5.1. Extending Electronic Document Delivery to Grey Files

What methods would be most appropriate for delivering documents from an archive to a user – FTP, FTAM, SMTP, X.400, etc? The choice is essentially between email and file transfer. The two most frequently used methods on Internet (and the related national academic networks) are email (generally, SMTP) to a mailbase host, or **Anonymous FTP**. Generally speaking, email is very convenient for ASCII-encoded files, such as text documents, but a file transfer protocol (such as FTP) is needed for binary data, such as fax-encoded page images – unless these are first converted to a 7-bit ASCII format.

The problem with **Anonymous FTP** is its high dependence on manual interactive control – though even this could be avoided if appropriate machine-driven interfaces were developed. To some extent, these already exist in the form of the WWW browser. Nevertheless, most present methods for access to Internet documents tend to be very labour-intensive. At each stage – manually finding, retrieving, and printing – the complexity and time-consuming nature of the procedures is a considerable disincentive to their wider use. In contrast, document delivery by more traditional means (namely, the photocopier and postal services) is very easy – just fill in the form and pick up the reply in the mail. The problem is the time delay between these two events.

Email (SMTP), in comparison with Anonymous FTP, is less complex to use but has suffered from the fact that many implementations cannot deal effectively with large binary files. Binary files must first be converted to 7-bit ASCII (and re-converted on delivery). Large files must be broken up into smaller chunks (then reassembled at the delivery site before printing). With later versions of X.400, however, neither of these issues is a problem. Similarly, the extension of SMTP to include the MIME protocol should overcome these difficulties as well. A mail-based document server, capable of dealing with a variety of document formats, including binary encoded images, then becomes an attractive possibility.

Over the past year, The British Library has been experimenting with the use of X.400 as a carrier for TIFF-encoded documents. The results have been very encouraging, at least as far as the technology is concerned. Documents can be scanned in at any acceptable resolution (normally 400 dpi from a flat-bed scanner, though higher rates are possible) and transferred as a TIFF file via email. On receipt at the requesting site, the message is automatically unbundled and the document sent to a page printer (using FTP as the transfer protocol for this stage), without the need for any manual intervention. Client notification of the document's arrival is also given by an automatically generated email message.

The principal advantage of using X.400 lies in the fact that the protocol itself handles questions of addressing, message queuing, error notification, and so on. With an FTP- (or FTAM-) based transfer, all this would have to be added on top. In addition, document request messages can be easily handled by the same email

channel, as can monitoring information, shipped messages, acknowledgements and billing.

In this context, the emerging vision for document delivery is a combination of the indexing information of *Inside Information* (or its RAPDOC equivalent) – extended to include electrostorage archives – coupled with email ordering and delivery. The inner workings of the system – whether based on FTP, FTAM or X.400 – should be invisible to the user. While Internet access methods are also moving in this direction, they still suffer from a burden of technical overload inherited from their "Unix expert" past.

5.2. Transfer Methods: Network Protocols

Beneath the two high-level document transfer facilities (ie, file transfer and email) lie the network protocols themselves. This is the realm of *open* versus *proprietary* networks, but also of *de facto* versus *de jure* standards.

A recent review pointed out that TCP/IP (generally considered as being the network protocol of choice for most academic computing services) is used in as much as 30% of all current (local and wide area) computer networks. OSI (which includes X.25), in comparison has only 3% of the market. The remaining 57% is split between various proprietary protocols, including IBM's SNA (30%) and DECnet.

The academic community in the Netherlands, as well as in the UK, are wedded almost exclusively to TCP/IP as the general means for wide area network access. In the case of the Dutch academic community, TCP/IP wide area access is provided over native IP, which is physically separate from the SURFnet X.25 network. A similar development is taking place in the UK, with native IP rapidly replacing IP over X.25. Internet is an international amalgamation of academic and commercial TCP/IP networks. It provides connections between, for example, the JANet network in the UK and SURFnet in the Netherlands. As such it is the natural carrier for document delivery services of the kind envisaged, including Grey Files.

Although FTP and SMTP are the native Internet protocols for file transfer and email, respectively, well-understood mechanisms are available for running X.400 email or FTAM file transfer (the OSI protocols) over TCP/IP networks. This means that one is not restricted to Internet protocols, even within a TCP/IP domain. The choice may then be made on the basis of functionality, performance, cost, and so on.

6. Document Presentation: Display and Printing

6.1. Display Facilities: General Issues

Printing and display probably cause more frustration than any other area of electronic document retrieval. Even when the material is in plain ASCII format there can be problems, such as mismatching page-lengths, for example. Nor does the move to PostScript for printed output completely resolve the problems, as variant

implementations can be incompatible. It remains to be seen whether the adoption of Acrobat by a significant number of commercial publishers will lead to an agreed standard that will overcome these difficulties.

Even where there are no technical problems, the need for a high degree of manual intervention creates frustration. For example, if the document has been delivered by email then the user must first remove the message header before sending it to a printer. Similar problems can occur with handling long filenames in transferring documents from archives to PCs via Anonymous FTP. For example, using multifile **get** and wildcard * commands to reduce the need for typing long filenames can come unstuck if more than one file is required.

In the case of TIFF-encoded documents, the technical difficulties are even greater. For a start, few printers are able to take a TIFF file and print it without it first being processed into some conformant format. This will often entail proprietary software or hardware, particularly if the printer is to be driven at anything near its optimal rate. Most solutions offered by the manufacturers of DIP equipment are of this kind.

Encapsulation of TIFF within PostScript is technically feasible, but results in a vastly expanded file size. This creates a significant delay on the link between computer and printer and slows the output considerably. In addition, the way in which PostScript handles images is very processor-intensive, which also slows output. The net result is that this method is not useable in any practical sense for TIFF-encoded documents of more than one page.

On Internet Anonymous FTP archives, TIFF files are invariably compressed. This requires the user to decompress the file with appropriate software prior to any attempt at printing. In general, this will mean still more manual intervention and usually entails a level of technical knowlege that would be outside the domain of most ordinary users. In other words, as presently organised, it is not the basis for a practical document delivery service. It leaves one longing for the simplicities of fax, or even the photocopier and post.

Nevertheless, the components and resources for dealing with all these problems do exist. It is primarily a matter of organising them in such a way that the process of *delivery to print* can operate automatically and transparently – without bothering the user with details of Lempel-Ziv compression and the like. That this has not happened is partly due to the nature of the present user community on Internet. Initiatives such as the IAFA, however, do represent a move in the right direction.

6.2. Document Browsers

Many of the problems associated with printing can be resolved by incorporating this function into a well-configured browser. As noted above, for use with the World Wide Web, numerous browsers for different platforms are readily available in the public domain. This has advantages for users, as there is a choice, while support for installing and maintaining browsers is free and readily available through the Internet.

If the browsing end of the chain is set up correctly, external viewers may be used to view special files, whether in the form of TEX-dvi files, PostScript files, GIFF images, MPEG movies, or TIFF images.

Printing must also be set up locally at the browsing end. Browsers offer the possibility to print hardcopy of files in a variety of formats (PostScript, ASCII), depending on local availability and system configuration.

For example, in the Mercury system, transactions between the Mercury client and different server databases are hidden from end users by a consistent user interface. All queries to any database in the system are made through the same (forms based) interface. The Mercury system leans heavily on the Z protocol for these database transactions.

The Mercury system offers possibilities to view (bit-mapped, TIFF) images of documents that are stored in a document base. These images could, in turn, serve as extra entry points for users of the Grey Files databases, so that where images of documents are available, they might get called up from inside Mercury without a user needing to hook up into the net.

6.3. Document Printing

In the experiments of the British Library, documents are transmitted in standard GEDI format (TIFF encoded images, with a tagged item header file). As mentioned earlier, the carrier is X.400 email. On receipt, the message is automatically unbundled to extract the header file (ASCII encoded) and the TIFF images. These are then further processed to convert them into the form required for an Oce/QMS 6500 image printer. This printer is one of the few on the market (and possibly the only one) with the ability to print directly from a TIFF file (with CCITT Gp4 compression). The printer can be set up on a local ethernet, with its own IP address. Files are then printed by the simple process of FTPing them from host to printer across the network. Even in this case, however, the original file must first be passed through a preprocessor *TIFF reader* which checks the format (resolution, image dimensions, etc.) and attaches a controlling header script to the file. *TIFF to screen* has many of the same problems, requiring special decoding and display software (and hardware).

One conclusion from these experiments has been that to duplicate the current document delivery services (based on photocopier and post) by use of the Internet as carrier, is not easy. Nor is it cheap (the Oce 6500 costs nearly £6K). Fax is undoubtedly a much easier (and far cheaper) option, which partly explains its adoption by CARL. The argument for using Internet is that, in the longer term, it should be more cost-effective and able to offer a wider range of services (colour images, higher resolutions, etc.). There is a fundamental issue here as to how an electronic document delivery service should be organised – fax machines, unfortunately, come with telephone numbers, not IP addresses.

In fact, a well-developed document delivery infrastructure based on the telephone and fax network already exists in the commercial world as part of the *audiotex* business. Fax-on-demand from voice- (or telephone keypad-) activated

document servers is a significant and rapidly growing market sector [Tuck, B. (1992b)].

The great difficulty here is the almost total decoupling of the fax network from the academic data networks (such as Internet). This is not likely to change in the short term. Deployment of broadband ISDN, based on technologies such as ATM (Asynchronous Transfer Mode) and service structures such as SMDS (Switched Multimegabit Data Service) may alter this factor in the longer term, though it remains very unclear how any real merger might come about.

In the meantime, document delivery services hover between the limited plain text forms now dominant in the network world and the low resolution images of fax-based services.

6.4. Printing Grey Files

It is anticipated that in the Grey Files project many of these problems would not appear. Since the project is aimed essentially at the electronic distribution of grey literature, copyright issues are avoided. This means that on-screen presentation of electronically stored material would be the normal mode of access. Hard copy *distribution* of print material does not therefore fall within the scope of the project and so needs only to be addressed as an aside.

Nevertheless, it is generally recognised that "there are no research memoranda that are not intended to be presented in print-on-paper form". Printing from the electronically stored (and displayed) grey file is therefore important. As mentioned earlier, this function would generally be incorporated into the document browser.

7. Summary and Conclusions

7.1. General Summary

By normal library standards the organisation of source documents on the Internet is haphazard and chaotic. Even in well-regulated services such as the UK *Mailbase* host, the indexing of material is rudimentary, the formats are extremely limited (no graphics) and while the methods of access are simple to use (basically, email file request) the process of generating printed output can be very laborious.

Poor response times on the network severely inhibit online interactive browsing. It is generally better to retrieve index files (again by using email request, or facilities such as Archie's `mail-to` option) for off-line browsing. This is particularly noticable for public access WAIS, Gopher or Archie systems. All those investigated during the course of this study exhibited response times that ranged from poor to apalling. Many were virtually unusable.

Library-based document request and delivery systems, such as *Inside Information*, suffer from the opposite problem. Access to high quality indexes is very easy (both online and offline) but requesting can be cumbersome compared to the Internet equivalents, and the time taken for the document to appear is excessive (several days, versus a few minutes, at most).

However speedy it may be, manual **Anonymous FTP** does not seem the way to go for a user-driven document supply server. It is too heavily orientated towards computer (or, more particularly, Unix) experts, and is very labour-intensive. The interactive advantages of Anonymous FTP are probably exagerated (and no longer necessary). The question of whether FTP is appropriate as a machine-driven protocol for automatic transfer is still open.

Fundamental problems still exist at the level of providing printed output of document images (eg. in TIFF format). Current network support for this format is poor. Likewise, available printer hardware tends to be proprietary and/or expensive. The same is generally true for screen-based TIFF readers. To provide the rapid decompression and high resolution display needed for adequate presentation is expensive and needs special hardware – though cheaper standard equipment may suffice for some applications.

In contrast, there is every likelihood that a standardised PDL (Page Description Language) based around extensions to PostScript (eg. Acrobat) could be adopted both by commercial publishers and printer manufacturers. This makes it feasible to build substantial databases of material in this format for general network distribution. *Commercial* electronic journals in this format are already being planned, so the impetus in this direction is quite high.

More general approaches to the encoding of material for network distribution are still at the formative stages. SGML is the obvious choice for a more generic approach, but there is as yet only limited familiarity with this within the academic community. Programmes such as the TEI (Text Encoding Initiative) may help to change this situation. In the short term, it is likely that a more ad hoc approach, based on familiar encodings such as PostScript and ASCII will have to be adopted.

At the same time, any initiative to create document image files will have to come from the library world. Current offerings on Internet seem to be largely restricted to test pictures and assorted trivia (Daffy Duck, Bart Simpson and sombody's girlfriend called Natalie, feature quite strongly). This indicates something of a cultural gap between the two domains. In spite of this, there would seem to be much to be gained by exploring the possibility of creating network accessible archives of *facsimile* documents, of both historical and current material – newspapers, music, manuscripts, etc. As an example, a substantial part of the Beowulf manuscript is now available in facsimile form from a server on the Internet, following a major initiative by the British Library. In general, collaboration with the IAFA working group on this issue would be advisable.

A working group under the IETF (the Networked Information Retrieval (NIR) Working Group) is currently evaluating all the emerging tools for network resource access (including WAIS, Gopher, Anonymous FTP, WWW, etc.) [NIR (1994)]. The NIR WG acts as a focus within the international academic network community for coordinating development of network access tools, including information servers.

Most of the text-only material that dominates current network sources should probably be considered in the class of non-print-orientated output. This would include newsgroups, bulletin boards and other informal publications. Such material

generally suffers from poor indexing, variable quality, and doubtful provenance, but nevertheless provides a valuable medium for informal research (and social) communication. Whether it should be accepted as *library material*, on the other hand, has to be questionable. The cost of providing adequate cataloguing and supporting access could be high.

In the longer term, evolution towards more formally produced *multi-media* (and, by implication, non-print) publications will create another new range of problems for libraries. This is material that from the beginning is designed to be played out on the computer screen rather than the printed page. In this respect it should perhaps be compared with film, video and sound recordings, for which special provision must be made.

7.2. Conclusions

Document origination

The existing document bases on the Internet are diverse in quality to say the least. The quality relates to the contents of the documents themselves as well as to the service. If a document server for Grey Files is to be viable, assurances to this effect will have to be made. The quality of service must be guaranteed by the institution; this means that

- the hardware will have to be up to measure for the task in order to offer acceptable response times;
- the server maintainer will have to be held responsible for the consistent availability of advertised documents;
- the local libraries will have to accept responsibility for adequate cataloguing;
- the quality of the contents of the documents will have to be guaranteed by the people responsible for the publication – in the case of Grey Files, this will be the editorial board of the series.

Encoding formats

The only reasonable encoding format at this stage appears to be PostScript.

- PostScript is widely available on a wide range of platforms, is extensively used in the academic world, and can be generated rather effortlessly by most commercially available wordprocessors.
- PostScript allows for indexing on the full text, which may prove to be a great asset.
- PostScript device drivers are available in the public domain.

Institutions that do not have PostScript capabilities will have to resort to publically available PostScript to ASCII conversion programmes.

In a later stage, after authors have become acquainted with SGML-based means of (network) publishing like the World Wide Web, some submissions will possibly be in the form of HTML source, and may contain links to other publications by the same or other authors, that reside on the same or other comparable document servers.

One development that should be watched in this context has to do with the evolution of HTML to a Text Encoding Initiative (TEI) compatible model [Hockey (1993); TEI (1993); Berners-Lee 1993b; Ragget (1993)].

Decisions about conversion and about allowing HTML grey files on the server should be made dependent on the outcome of the discussion mentioned above.

Using SGML as an encoding format is – at this stage – not advisable: there is not enough support for SGML to warrant the trouble of instructing users to standardise on this format.

Because of the fact that Adobe's Acrobat is both expensive and not yet widely available in the mainstream of computer-using academia – our primary audience – it should not be used in the pilot.

On the other hand, developments around Acrobat certainly form an area that deserves full attention. Not in the last place because it supposedly allows the extraction of document contents *and* document structure from one and the same file, the Portable Document Format (PDF) file [see Adobe (1993) for details].

ASCII-encoded (plain text) versions of the papers should possibly also be made available. The reason for this is that previewing PostScript files is not possible on character based terminals – and there are still a large number of these around

Information retrieval and document requests

The main problem in this area is *poor indexing*. One of the options that seem promising is the idea of an extended archie server. Alternatively, and this is recommended, litterature of this sort should be catalogued analogously to their non-electronic pendants, so that searching a central or national bibliographical database may yield physical as well as electronic locations.

There is one development in the area of naming electronic resources that need close attention during the pilot phase. One is the evolution of a new naming scheme, the Uniform Resource Name (URN), that aims at providing a more persistent naming mechanism for network information resources, and which includes formal bodies of authority that allocate URNs and watch and monitor their use [Theise (1993); Weider — Deutsch (1993)].

Additionally, it is desirable that document banks such as Grey Files should also be accessible and searchable with WAIS.

Document delivery methods

Although a protocol like manual **anonymous ftp** suffers from 'a burden of technical overload inherited from the "Unix expert" past', it seems that, once hidden from the user under a user-friendly layer, it would be very suitable for our purpose. The layer that seems most promising in this respect is the World Wide Web.

The choice for WWW as the vehicle for document delivery stems from the following observations:

Flexibility The rich yet straightforward design of World Wide Web with its clearly separable components (HTML markup, URL locators and HTTP protocol) means that it is a flexible basis on which to develop distributed multimedia applications.

Hyperlinks are very well supported in the World Wide Web, and offer extra entry points to Grey Files – apart from PICA's GGC and other (special interest) reference databases (Attent, OPC).

Integrated solution Because WAIS, GOPHER and anonymous FTP servers may all be accessed from Web clients, WWW serves as an important integrating tool for information services.

Penetration and growth Although GOPHER surpasses the World Wide Web in the number of servers available, the rate of growth in WWW usage is greater than that of GOPHER. There seems to be a growing realization in the community that GOPHER is over-simplistic for many purposes.

Standardization The URL specification has already been published as an Internet draft, and has been adopted as an important component of the proposed Internet integrated information infrastructure. The current version of HTML is about to follow suit. The use of SGML as the basis of HTML complies with the percieved importance of SGML for hypermedia and for the dissemination of documents in electronic form.

Software status CERN has recently placed the WWW code in the public domain. This is unlike all the other candidate technologies, which all have restrictions of one form or another. In the case of GOPHER, these restrictions are already causing commercial users to look elsewhere.

Document presentation

The components and resources for dealing with decompression, decoding and on-line browsing of electronic documents are available, especially with regard to PostScript and a number of other representation formats (GIF, TIFF, MPEG). Document browsers and viewers are available in the public domain for different platforms, so that the user has a choice.

Printing may be an issue, as PostScript printers, even though widely available, may not be accessible everywhere. In some cases, importing the PostScript file in a wordprocessor, and exporting it in another format may be a solution.

8. Recommendations

There is little doubt that World Wide Web has become the most popular system for integrating information sources on the Internet. Currently available alternatives – such as Gopher, Archie, WAIS, or X.500 – lack much of its flexibility and other positive characteristics. For this reason, the proposal outlined in the Appendix to develop a demonstrator document server based on WWW is recommended.

To provide a significant evaluation scenario it would clearly need to involve more than one server. Ideally, it might be three or more, each holding material in similar academic subjects (to allow for hyperlinks) and with wide-area connections to evaluate network requirements. Servers at Tilburg and Maastricht, each holding documents on economics, for example, would seem to be ideal.

The recommended actions include the creation of a significant demonstrator of WWW and to link the contents of this into the RAPDOC document request and delivery system. The use of WWW as a general interface to RAPDOC services, on the other hand, needs further examination before any serious comitment should be made.

Participation of the UK in such experiments might be encouraged by setting up a WWW server in parallel with those at Tilburg and Limburg (with documents on a common subject). There could also be some interest in extending the DDX experiment to allow the ordering (by X.400) of material from WWW document bases. The British Library is already implementing such a server, based on its own collection of internal publications (information brochures, rather than serials, or grey files). PC browsers (Cello and Mosaic) are being used, as well as the X-windows versions. They are also funding a joint project in electronic publishing with the Institute of Physics, which will use WWW as one of the access mechanisms for the publications. There are believed to be a number of similar experiments at other sites within the UK and the Netherlands. This activity will certainly grow over the next year or so, and with the inauguration of SuperJANET and other advanced networks, the interlinking of WWW servers over high-speed circuits will provide interesting experimental data on this new form of publishing.

In summary, the recommendations from this study are the following:

1. Set up a Grey Files WWW server as a demonstrator project
2. Encourage a similar initiative within the UK
3. Investigate further the possibility of using WWW as an access point for document ordering
4. Investigate email access to the Grey Files database and its integration with RAPDOC and DDX
5. Investigate the additional possibilities that might come about through linking WWW document servers over high-speed networks such as SuperJANET

References

Adobe Systems Incorporated. (1993) *Portable Document Format Reference Manual.* ISBN 0-201-62628-4. Addison-Wesley Publishing Company. Reading, Massachussets.

Berners-Lee, T. (1993a) *Uniform Resource Locators.* Internet draft. IETF URL Working Group. CERN, Geneva (CH).

Berners-Lee, T. (1993b) *Hypertext Markup Language.* Internet Draft. URL: http://info.cern.ch/hypertext/WWW/MarkUp/MarkUp.html

Berners-Lee, T. (1993c) *Protocol for the Retrieval and Manipulation of Textual and Hypermedia Information.* Internet Draft. URL: http://info.cern.ch/hypertext/WWW/Protocols/HTTP/HTTP2.html

Deutsch — Emtage: *Publishing Information on the Internet with Anonymous FTP*

GEDI, (1991) *Electronic Document Delivery: Towards further standardization of international interchange*, Proposals of the Group on Electronic Document Interchange (GEDI), October 1991

NIR (1994) *Status Report on NIR Tools*, available by Anonymous FTP from mailbase.uk.ac in directory pub/lists/nir/files – filename nir.status.report.].

Hockey, S. (1993) *The ACH-ACL-ALLC Text Encoding Initiative: An Overview.* June 1991 (Rev. Feb 28, 1993) Document Number: TEI J16. Available from LISTSERV@UICVM.CC.UIC.EDU.

TEI (1993) *Notes from WWW/TEI Meeting.* URL: http://curia.ucc.ie/info/net/WWW-TEI_Meeting_Notes.html

Theise, E. S. (1993) *Curling up to Universal Resource Locators.* URL: gopher://gopher.well.sf.ca.us/00/matrix/internet/curling.up.02

Tuck, B. (1992a) *A Research Agenda for Scientific and Technical Information*, AGARD, AR-316, November 1992

Tuck, B. (1992b) *Fax-on-demand.* Electronic Documents; Vol. 2.

Ragget, D. (1993) *HTML+ (Hypertext Markup format).* Internet Draft. IETF. Hewlett Packard Laboratories.

Weider, C. and P. Deutsch (1993) *Uniform Resource Names.* Internet Draft. IETF URI Working Group. Bunyip Information Systems, Montreal (CDA).

OVERVIEW OF ITK RESEARCH REPORTS

No	Author	Title
1	H.C. Bunt	On-line Interpretation in Speech Understanding and Dialogue Sytems
2	P.A. Flach	Concept Learning from Examples Theoretical Foundations
3	O. De Troyer	RIDL*: A Tool for the Computer-Assisted Engineering of Large Databases in the Presence of Integrity Constraints
4	M. Kammler and E. Thijsse	Something you might want to know about "wanting to know"
5	H.C. Bunt	A Model-theoretic Approach to Multi-Database Knowledge Representation
6	E.J. v.d. Linden	Lambek theorem proving and feature unification
7	H.C. Bunt	DPSG and its use in sentence generation from meaning representations
8	R. Berndsen and H. Daniels	Qualitative Economics in Prolog
9	P.A. Flach	A simple concept learner and its implementation
10	P.A. Flach	Second-order inductive learning
11	E. Thijsse	Partical logic and modal logic: a systematic survey
12	F. Dols	The Representation of Definite Description
13	R.J. Beun	The recognition of Declarative Questions in Information Dialogues
14	H.C. Bunt	Language Understanding by Computer: Developments on the Theoretical Side
15	H.C. Bunt	DIT Dynamic Interpretation in Text and dialogue
16	R. Ahn and H.P. Kolb	Discourse Representation meets Constructive Mathematics

No	Author	Title
17	G. Minnen and E.J. v.d. Linden	Algorithmen for generation in lambek theorem proving
18	H.C. Bunt	DPSG and its use in parsing
19	H.P. Kolb and C. Thiersch	Levels and Empty? Categories in a Principles and Parameters Approach to Parsing
20	H.C. Bunt	Modular Incremental Modelling Belief and Intention
21	F. Dols	Compositional Dialogue Referents in Prase Structure Grammar
22	F. Dols	Pragmatics of Postdeterminers, Non-restrictive Modifiers and WH-phrases
23	P.A. Flach	Inductive characterisation of database relations
24	E. Thijsse	Definability in partial logic: the propositional part
25	H. Weigand	Modelling Documents
26	O. De Troyer	Object Oriented methods in data engineering
27	O. De Troyer	The O-O Binary Relationship Model
28	E. Thijsse	On total awareness logics
29	E. Aarts	Recognition for Acyclic Context Sensitive Grammars is NP-complete
30	P.A. Flach	The role of explanations in inductive learning
31	W. Daelemans, K. De Smedt and J. de Graaf	Default inheritance in an object-oriented representation of linguistic categories
32	E. Bertino and H. Weigand	An Approach to Authorization Modeling in Object-Oriented Database Systems
33	D.M.W. Powers	Multi-Modal Modelling with Multi-Module Mechanisms: Autonomy in a Computational Model of Language

No	Author	Title
34	R. Muskens	Anaphora and the Logic of Change*
35	R. Muskens	Tense and the Logic of Change
36	E.J. v.d. Linden	Incremental Processing and the Hierarchical Lexicon
37	E.J. v.d. Linden	Idioms, non-literal language and knowledge representation 1
38	W. Daelemans and A. v.d. Bosch	Generalization Performance of Backpropagation Learning on a Syllabification Task
39	H. Paijmans	Comparing IR-Systems: CLARIT and TOPIC
40	R. Muskens	Logical Omniscience and Classical Logic
41	P. Flach	A model of induction
42	A. v.d. Bosch and W. Daelemans	Data-oriented Methods for Grapheme-to-Phoneme Conversion
43	W. Daelemans, S. Gillis, G. Durieux and A. van den Bosch	Learnability and Markedness in Data-Driven Acquisition of Stress
44	J. Heemskerk	A Probabilistic Context-free Grammar for Disambiguation in Morphological Parsing
45	J. Heemskerk and A. Nunn	Dutch letter-to-sound conversion, using a morpheme lexicon and linguistic rules
46	A. HH. Ngu, R. Meersman and H. Weigand	Specification and verification of communication constraints for interoperable transactions
47	J. Jaspars and E. Thijsse	Fundamentals of Partial Modal Logic
48	E. Krahmer	Partial Dynamic Predicate Logic
49	W. Daelemans	Memory-Based Lexical Acquisition and Processing
50	G. Rentier	A Lexicalist Approach to Dutch Cross Serial Dependencies
51	R. Muskens	Categorial Grammar and Discourse Representation Theory

No	Author	Title
52	M. Rats	Topic-Comment Structures in Information Dialogues
53	G. Rentier	Extraction in Dutch with Lexical Rules
54	H. Bunt	Dialogue control functions and interaction design
55	A. van Horck	Document Servers and the Electronic Distrubution of Grey Literature

ITK: P.O. BOX 90153 5000 LE TILBURG THE NETHERLANDS